Future Generation Computer Systems 57 (2016) 56–76

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

A model to compare cloud and non-cloud storage of Big Data

Victor Chang [a,*], Gary Wills [b]

[a] School of Computing, Creative Technologies and Engineering, Leeds Beckett University, Leeds, UK
[b] School of Electronics and Computer Science, University of Southampton, Southampton, UK

HIGHLIGHTS

- Organizational sustainability modeling (OSM) compares Cloud and non-Cloud storage.
- We identify factors affect performance and design ways to make fair comparisons.
- We explain how to use OSM including its definitions, input and output.
- We present two case studies of Big Data storage with 40 runs to support.
- Results are analyzed and presented with data analysis and visualization.

ARTICLE INFO

Article history:
Received 11 July 2015
Received in revised form
28 August 2015
Accepted 6 October 2015
Available online 26 October 2015

Keywords:
Organizational sustainability modeling (OSM)
Comparison between Cloud and non-Cloud storage platforms
Real Cloud case studies
Data analysis and visualization

ABSTRACT

When comparing Cloud and non-Cloud Storage it can be difficult to ensure that the comparison is fair. In this paper we examine the process of setting up such a comparison and the metric used. Performance comparisons on Cloud and non-Cloud systems, deployed for biomedical scientists, have been conducted to identify improvements of efficiency and performance. Prior to the experiments, network latency, file size and job failures were identified as factors which degrade performance and experiments were conducted to understand their impacts. Organizational Sustainability Modeling (OSM) is used before, during and after the experiments to ensure fair comparisons are achieved. OSM defines the actual and expected execution time, risk control rates and is used to understand key outputs related to both Cloud and non-Cloud experiments. Forty experiments on both Cloud and non-Cloud systems were undertaken with two case studies. The first case study was focused on transferring and backing up 10,000 files of 1 GB each and the second case study was focused on transferring and backing up 1000 files 10 GB each. Results showed that first, the actual and expected execution time on the Cloud was lower than on the non-Cloud system. Second, there was more than 99% consistency between the actual and expected execution time on the Cloud while no comparable consistency was found on the non-Cloud system. Third, the improvement in efficiency was higher on the Cloud than the non-Cloud. OSM is the metric used to analyze the collected data and provided synthesis and insights to the data analysis and visualization of the two case studies.

1. Introduction

Cloud Computing is being adopted and investigated by an increasing number of organizations to demonstrate proofs-of-concepts and successful adoption. In the process of adopting and using Cloud Computing services, masses of data from the people (users and stakeholders) and projects (experiments, simulations, images and documents) have been produced, exchanged and stored. As a result, sophisticated techniques are required to deal with increasing demands for data processing, management and analytics [1–3]. Big Data has five characteristics: volume, velocity, variety, veracity and value [4]. Volume refers to the size of the data for processing and analysis. Velocity refers to the rate of the data growth and usage. Variety means the different types and formats of the data used for processing and analysis. Veracity concerns the accuracy of results and analysis of the data. Value is the added value and contribution offered by data processing and analysis. Due to the maturity of Cloud technologies and demands in the use of data, the storage of Big Data is an important topic in Cloud research. Maturity of technologies includes the readiness of Web 2.0, virtualization, data center technologies, fast network speeds and bandwidths, libraries and APIs for Cloud Computing. MapReduce is a popular framework adopted by Cloud Computing

* Corresponding author.
E-mail address: V.I.Chang@leedsbeckett.ac.uk (V. Chang).

http://dx.doi.org/10.1016/j.future.2015.10.003
0167-739X/

to process and analyze data. It splits into map and reduce functions, whereby "maps" categorizes the same types of data together and "reduces" then performs the processing of the data to generate the outputs. Often additional algorithms have to be written to ensure smooth processing and transition in the data processing. For example, an optimize function can be written to accelerate the processing time and a visualize function can transform numerical outputs so that users without much technical knowledge can understand the outputs more easily [5].

Big Data in the Cloud offers opportunities for scientists in providing a faster and more accurate technique to analyze their experimental data. At the end of each experiment, terabytes of data can be generated ranging from numerical outputs, the scientific calculations, documentation, images of all kinds (DNAs, tumor and proteins) to datasets, both raw and processed. This will require excellent data processing and management strategies and policies in place, with both automated and manual processing as well as monitoring systems to ensure Big Data services in the Cloud can run smoothly and minimize discrepancies such as fluctuation in network performance, execution time, and termination of services due to job failures. The literature suggests that scientists have used public Clouds to process large scale experiments [4,6,7]. However, sensitive data such as patients' records and body images such as tumor and surgery related information, should not be in public domains. All these data should only be within the hospital and not in any public clouds. Hence, the design and implementation of private clouds is essential for biomedical scientists to generate, process, update, archive and store their data. This paper will describe private cloud development for biomedical scientists, whereby high-performance Cloud storage and Big Data processing can be achieved. Our research contributions include:

- Direct comparisons between Cloud and non-Cloud platforms about their backup performance.
- A model to calculate improvement in efficiency of Cloud systems over non-Cloud systems for biomedical data backup.
- Data analysis and visualization.

The breakdown of this paper is as follows. Section 2 describes the related literature. Section 3 explains the system design and implementation. Section 4 presents the OSM model as the metrics for these experiments. Section 5 examines what control measures were in place to ensure an equitable comparison of the non-Cloud and Cloud based back-up systems. Section 6 presents the results of the experiments. Section 7 presents a brief discussion and Section 8 sums up the paper with the conclusion and future work.

2. Related work

The list of selected literature starts with backgrounds, the process of getting popularity and explanations about the problems associated with the models proposed by the following authors.

Calero and Aguado [8] propose architectures for monitoring Cloud Computing infrastructures and explain their internal and external approaches for monitoring physical and virtual machines. They present monitoring VMs from Cloud consumers point of view and architectures for monitoring in the Cloud. Their approaches are on the full management and monitoring of VMs and performance but do not provide remedies when network outage or latency causes performance downgrade.

Calheiros et al. [9] develop their ARIMA-based predictor for provisioning of virtual instances and only focus on the short term predictions and short-term impact in their QoS and SaaS application. Additionally, their evaluation is based on four-weeks of a single web workload trace.

Bossche et al. [10] focus on IaaS optimization with load prediction. They develop their algorithms based on ARIMA, Holt–Winters and exponential smoothing techniques to achieve renewal contract policies and load prediction. Instead of doing one web log experiment like [9], they adopt 51 real world web application load traces to evaluate their performance although their approaches are not monitoring live systems or applications in real time.

Bower et al. [11] propose their high-availability and integrity layer (HAIL) for Cloud Storage. They use mathematical proof and experiments to validate HAIL. In the domain of Big Data in the Cloud, experiments should focus on transferring data across different Clouds. Their results on availability are insightful but they do not have results for the total time taken, failure rate and performance downgrade caused by latency and large size of files.

Wang et al. [12] propose a framework of workload balancing and resource management for Cloud Storage known as "Swift". They use Swift to discover overloaded nodes and under-loaded nodes in the cluster and then try to make a good balance in all the nodes. A better alternative would be to balance the workload distribution before starting the experiments.

Rahman and Rahman [13] propose a Capital Asset Pricing Model (CAPM) for Grid Computing for e-negotiation and resource allocation. However, they do not have continuous monitoring systems or detailed experimental results on data transfer, failure rates and issues caused by latency.

Latch et al. [14] also use relative performance to present their Bayesian clustering software and their key performance indicators are presented as the percentages of improvement. Their work on relative performance needs to be leveraged and adopted by real case studies. Relative performance is defined as the improvement in performance between the old and new service and often the expected outcome is that there is an improvement after adopting new services such as a Cloud Computing service.

The selected literature represents idea and systems that have areas of merits, however, their insufficiencies help focus our research. In as much as; none of the proposed models have investigated performance between a Cloud and non-Cloud system, or how to analyze the data from the Cloud and non-Cloud system. This system should demonstrate Big Data in the Cloud, having experiments on transferring data from one place to another and have the Cloud Storage capacity to offer such a set of services. Our metric (the OSM model) can be instrumental to analyze data, represent the outputs so that the meanings can easily be understood by the stakeholders and system managers, something that is often implicit in the data and difficult for not experts to understand.

3. System design

This paper describes a real case study in which a new Cloud was designed for biomedical scientists who were required to back up large amounts of data. The new Cloud based back-up service is fast and reliable. We will firstly present the old system (non-Cloud) and new Cloud, based service for a National Health Service (NHS) Trust in the UK. The NHS Trust has invested in the Cloud based service to ensure that all data can be backed-up safely on their systems. The Cloud based service was required to undertake the back-ups, while allowing scientists to carry on with their research and development that produces data that was to be stored safely.

The NHS Trust involved include Guy's and St. Thomas' NHS Trust (GSTT) and King's College London (KCL). A Storage Area Network (SAN) was set up in an IT hub located at St. Thomas' Hospital in 2007 for scientists based at Guy's Hospital. The scientists were involved in cancer research (specifically breast cancer) and they produced hundreds of images and data records after each surgery, experiment or simulation. Backup files included data records about patients such as medical records and tumors, detailed descriptions,